D464–D469 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 26 November 2012*
doi:10.1093/nar/gks1166

ChemProt-2.0: visual navigation in a disease chemical biology database

Sonny Kim Kjærulff[1], Louis Wich[1], Jens Kringelum[1], Ulrik P. Jacobsen[1], Irene Kouskoumvekaki[1], Karine Audouze[1], Ole Lund[1], Søren Brunak[1], Tudor I. Oprea[1,2] and Olivier Taboureau[1,3,*]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark, [2]Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131-0001, USA and [3]Department of Life Sciences, UMR-S973, MTi, University Paris Diderot, F-75013 Paris, France

Received September 14, 2012; Revised October 26, 2012; Accepted October 28, 2012

ABSTRACT

ChemProt-2.0 (http://www.cbs.dtu.dk/services/ChemProt-2.0) is a public available compilation of multiple chemical–protein annotation resources integrated with diseases and clinical outcomes information. The database has been updated to >1.15 million compounds with 5.32 millions bioactivity measurements for 15 290 proteins. Each protein is linked to quality-scored human protein–protein interactions data based on more than half a million interactions, for studying diseases and biological outcomes (diseases, pathways and GO terms) through protein complexes. In ChemProt-2.0, therapeutic effects as well as adverse drug reactions have been integrated allowing for suggesting proteins associated to clinical outcomes. New chemical structure fingerprints were computed based on the similarity ensemble approach. Protein sequence similarity search was also integrated to evaluate the promiscuity of proteins, which can help in the prediction of off-target effects. Finally, the database was integrated into a visual interface that enables navigation of the pharmacological space for small molecules. Filtering options were included in order to facilitate and to guide dynamic search of specific queries.

INTRODUCTION

In recent years, there has been a shift from the traditionally secret experimental data kept by the pharmaceutical industry to a more open-access culture in relation to data sharing (1). For this reason, we have been witnessing a steady increase in public repositories of bioactive small molecules such as ChEMBL (2) and PubChem (3). However, as public repositories of bioactive small molecules have only just recently been made available, the problem of how to handle chemical entities is still largely unsolved. Pooling data from small molecule databases poses special problems. Even though standards have been widely adopted to describe genes and proteins (e.g. Ensembl ID, Entrez ID for genes, and UniProt ID for proteins), small molecule identifiers, as well as measures for properties such as biological activities, are not necessarily standardized across different resources (4).

One could claim that the bottleneck in understanding how small molecules perturb biological systems is no longer in the generation, gathering and availability of experimental data but in their organization, presentation and visualization; in other words, in the development of centralized systems that would better enable their exploitation. The problem is not only how to extract data from different (federated) resources, it is also important to provide solutions that facilitate provenance tracking, visualization, uniform and systematic description of data and their integration in ways that can preserve the semantic relationships between the different entities.

Furthermore, the number of failures of drug candidates in advanced stages of clinical trials has increased and the number of submissions for US Food and Drug Administration (FDA) approval has decreased in the last decade. One of the reasons may be our reductionist approach to discovery, whereby a complex system, namely a drug and its metabolites interacting with many proteins across multiple cellular compartments and tissues over time, is reduced to a simplistic ligand–target interaction model. This is probably too crude and emphasizes the

*To whom correspondence should be addressed. Tel: +45 4525 2489; Fax: +45 4593 1585; Email: otab@cbs.dtu.dk

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

need to look at the effects of compounds on global systems aided by the integration of multiple biological and temporal data sources.

With the emerging fields of chemogenomics (5), systems pharmacology (6) and systems chemical biology (7,8), it becomes feasible to investigate the drug action at different levels from molecular to pathway, cellular, tissues and clinical outcomes (9). For example, it has become apparent that many common diseases such as cancer, cardiovascular diseases and mental disorders are much more complex than initially anticipated, as they are caused by multiple molecular and cellular dysfunctions rather than being the result of a single defect. Therefore, network-centric therapeutic approaches that consider entire pathways rather than single proteins must be investigated (10).

Among the recent advances in the field of systems chemical biology, servers supporting drug profiling such as STITCH (11), DisGENET (12) or the new database PROMISCUOUS (13) should be mentioned. STITCH3 provides confidence scores that reflect the level of confidence and significance of compound–protein interactions. PROMISCUOUS is a resource focused on drug compounds, including withdrawn and experimental, containing drug–protein interaction and side-effect (SE) information. DisGENET is a comprehensive gene–disease association database focused on the current knowledge of human genetic diseases including Mendelian, complex and environmental diseases.

We have previously reported the development of ChemProt, a disease chemical biology database (14). Compared with other approaches, ChemProt-1.0 offered a high level of integration of chemical and biological data, including internally curated disease-associated protein–protein interactions (PPIs) (15). Here, we present the second release of ChemProt, a resource of annotated and predicted disease chemical biology interactions. ChemProt-2.0 can be accessed at http://www.cbs.dtu.dk/services/ChemProt-2.0/. The present release contains a compilation of over 1 100 000 unique chemicals with biological activity for >15 000 proteins. We have added a visual interface that supports user-friendly navigation through the data, biological activities and disease associations. ChemProt-2.0 now enables the user to query the database not solely by chemicals or proteins but also through therapeutic effects, adverse drug reactions and diseases. The similarity ensemble approach (SEA) developed by Keiser *et al.* (16) has also been implemented, so that protein sequence similarity can be used when examining chemical promiscuity. With these updates, ChemProt-2.0 offers an integrative approach to understand the impact of small molecules on biological systems and contributes to the investigation of molecular mechanisms related to diseases and clinical outcomes. A workflow of the implementation is shown in Figure 1.

DATA SOURCES

Chemical–protein interactions data were gathered in June 2012 from updated open-source databases ChEMBL

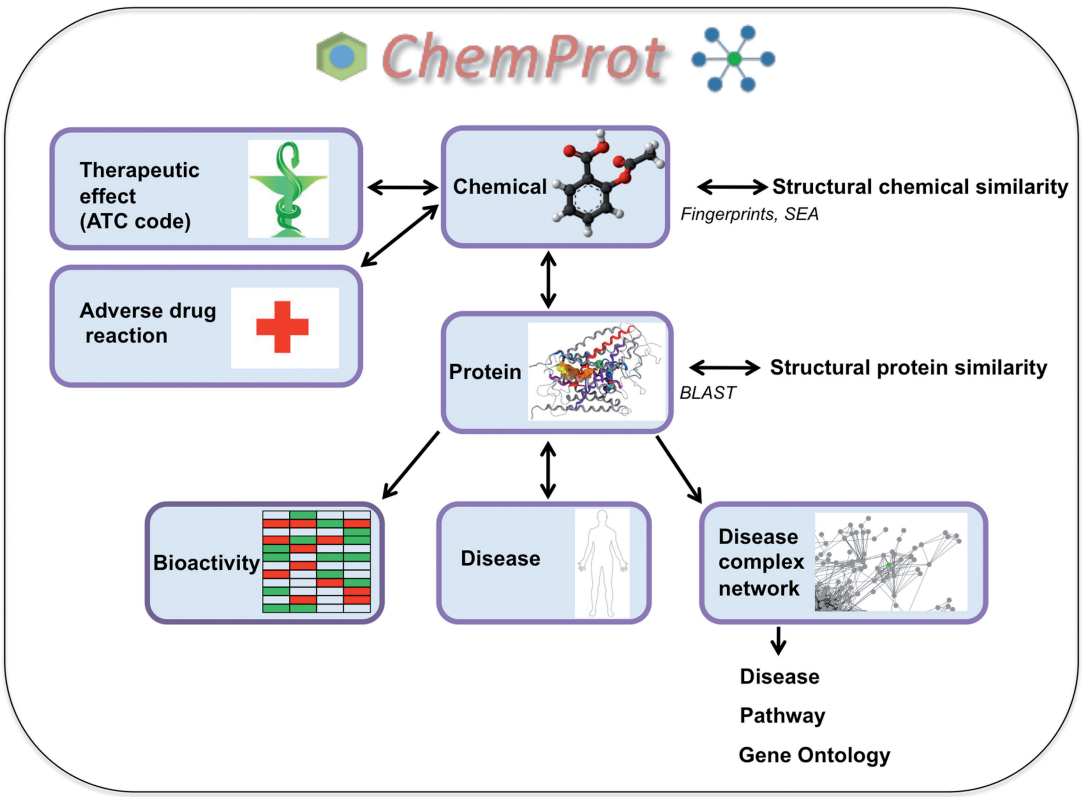

Figure 1. A workflow of the functionalities in ChemProt-2.0 is depicted. User can query ChemProt-2.0 using chemical, protein, disease, ATC code and SEs. Outcomes from the query are represented with the arrows.

(version 14), BindingDB (17), PDSP Ki database (18), DrugBank (version 3.0) (19), PharmGKB (20), active compounds from the PubChem bioassay (2012) targeting human proteins and the two commercial databases: WOMBAT (version 2011) and WOMBAT-PK (version 2011) (21). The IUPHAR-DB database (22) was also integrated in the new version of ChemProt-2.0. Chemical–protein annotations that lack explicit bioactivity data might be of interest in the mining of a large and diverse integrated database. Therefore, we included also data from CTD (23) and STITCH (11). CTD extracts literature data about environmental chemicals and how they modulate gene expression, whereas STITCH provides chemical–protein relationships from text mining the co-occurrence of a chemical term and a protein (gene) term in MEDLINE abstracts. Clinical outcomes were of special interest in this version and we decided to include information from the Anatomical Therapeutic Chemical (ATC) Classification System (24) developed by the World Health Organization, as well as SE data from Dailymed (http://dailymed.nlm.nih.gov/dailymed/).

From a biological perspective, we updated our internal human interactome platform to reach 14 421 genes interacting through 507 142 unique PPIs. The updated version of OMIM (25), GeneCards (26), KEGG (27), Reactome (28) and Gene Ontology (29) databases was also downloaded (June 2012), curated and integrated in ChemProt-2.0. Also, the human disease network developed by Goh *et al.* (30) was integrated, allowing association of proteins to disease categories.

PREDICTIONS AND METHODS

Based on the assumption that compounds sharing similar structure have potential similar bioactivities, we encoded the chemical structure with two different types of fingerprints: the 166 MACCS key which encode the presence or absence of some predefined substructural or functional groups (31) and the FP2 fingerprints computed with OpenBABEL (32). Chemical similarity between two compounds is quantitatively assessed using the Tanimoto coefficient. By including the SEA method (16), one can also predict potential new targets for a compound. For the internal development of SEA, compounds with an activity value <100 μM were considered (only IC50, EC50, Potency, AC50, Ki values were used). Furthermore, to complete the set of active protein ligands, annotated compound–protein interactions from CTD, DrugBank and PharmGKB were also included, together with annotated protein–compound in the STITCH database. For this dataset, the raw similarity score, i.e. the sum of ligand pair wise Tanimoto coefficients based on the FP2 fingerprint, is 0.44. All proteins with more than five bioactive ligands were considered.

In addition, for all protein targets, we operated under the assumption of promiscuity, i.e. proteins with high-sequence similarity may share similar functions and may be targeted by the same compound (likely with different bioactivities). Protein sequences were obtained from Uniprot (33), and sequence comparisons were computed using BLASTP (34). The similarity of two sequences was assessed using an *E*-score, an expectation value related to the probability that sequence similarity between two proteins is not achieved by random chance (34). We filtered the output and proteins with an *E*-value $<10^{-10}$ (as default) are depicted.

With respect to SEs, 988 small molecule drugs were matched against 174 SE as described (35). Term frequency vectors compiled from Dailymed were integrated in ChemProt-2.0 and proteins associated to each drug are then depicted.

VISUAL INTERFACE

In ChemProt-2.0, a visual interface was implemented to facilitate the visualization of the results using HTML 5 and JavaScript. The core of the interface has been designed in the form of a heatmap. The chemical–protein associations are depicted in a pie-chart heatmap where each pie corresponds to the database from which we gathered the information. Hovering over the pie-charts with the pointer, activity values are then displayed. The user can select different display settings (circles, fill and rectangles). A valuable feature is the handling of multiple activities that have been gathered for a given compound–target pair by selecting 'All' values. A color spectrum from blue (low activity) to red (strong activity) is used to indicate the activity (Figure 2). It is also possible to select a specific database or/and a specific activity type and define a range of activities (threshold) of interest in order to optimize the query. Results from the SEA approach are also integrated in the 'Activity Type'.

The compound query is always shown in the first column followed by similar compounds (sorted in descending order of similarity) whereas the protein queried is depicted in the first row. To optimize the display, the heatmap is limited to a section of 100 rows × 100 columns. If the chemical–protein matrix is larger, we have included an arrow feature (→) that allows the user to upload the next 100 data items for both axes. The user has still the possibility to view the data in a table format and to download the results in a flat-file format. In the table format, display mode the user can dynamically sort and group the activities according to compound, target, species, activity type, etc.

A second heatmap that depicts protein–disease categories is also integrated, which suggests proteins that may be involved in diseases. Next to it, the 'Diseases' link redirects the user to the disease-associated proteins complex around the selected protein. A new, dynamic interface has been implemented, where the proteins associated to a biological term are shown when highlighting the term of interest (Figure 3).

APPLICATIONS

The ChemProt-2.0 database interface is accessible freely online. In addition to the chemical and protein search that was previously implemented, the user can search by diseases, ATC codes and SEs. For example, the query 'epilepsy' returns 2662 compounds active on 13 proteins associated to this disease. Similarly, looking for the SE 'hallucinations', 15 drugs (with the term frequency

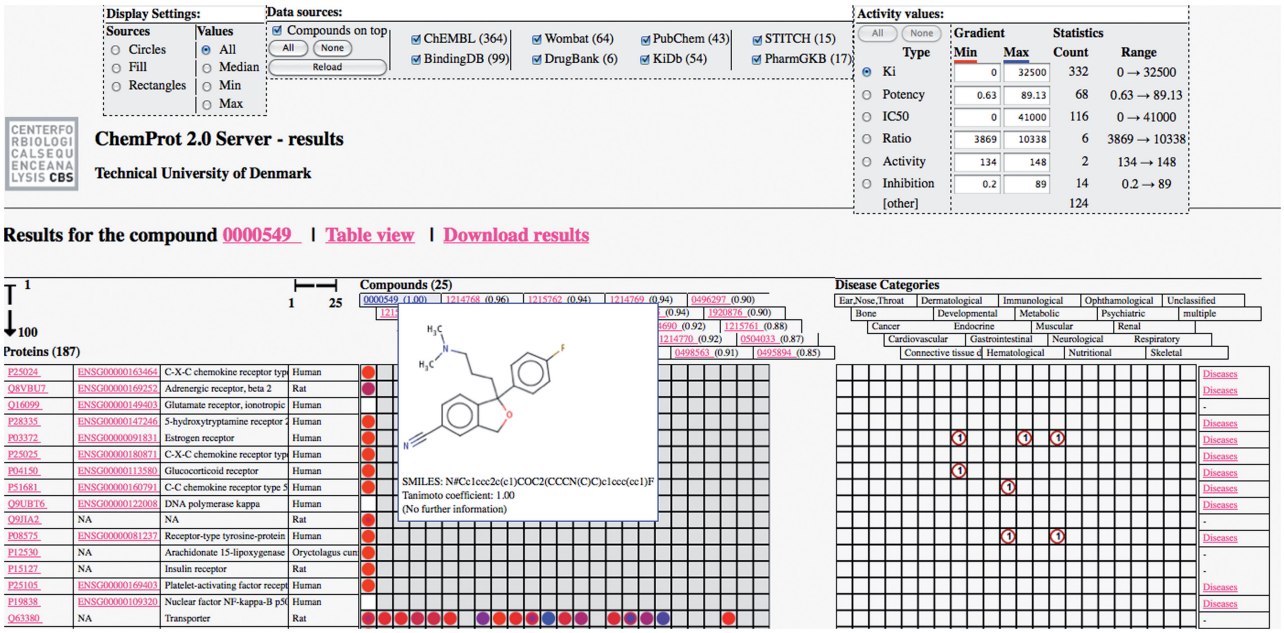

Figure 2. Example of the graphical interface output based on a compound query. On the top, user can specify the query using the display settings. The heatmap on the left represents the bioactivities gathered for the input compound (in blue) and structurally similar compounds (in pink) in the *X*-axis and the proteins in the *Y*-axis. A color spectrum from blue (low) to red (high) is used to represent the activity. If several binding data have been measured for the same chemical–protein interaction, intensity of the colors is represented inside the circle. It is shown for example for the dopamine transporter (Q63380). The heatmap on the right describes the disease categories annotated to a protein. The value inside the circle represents the number of diseases associated to a protein.

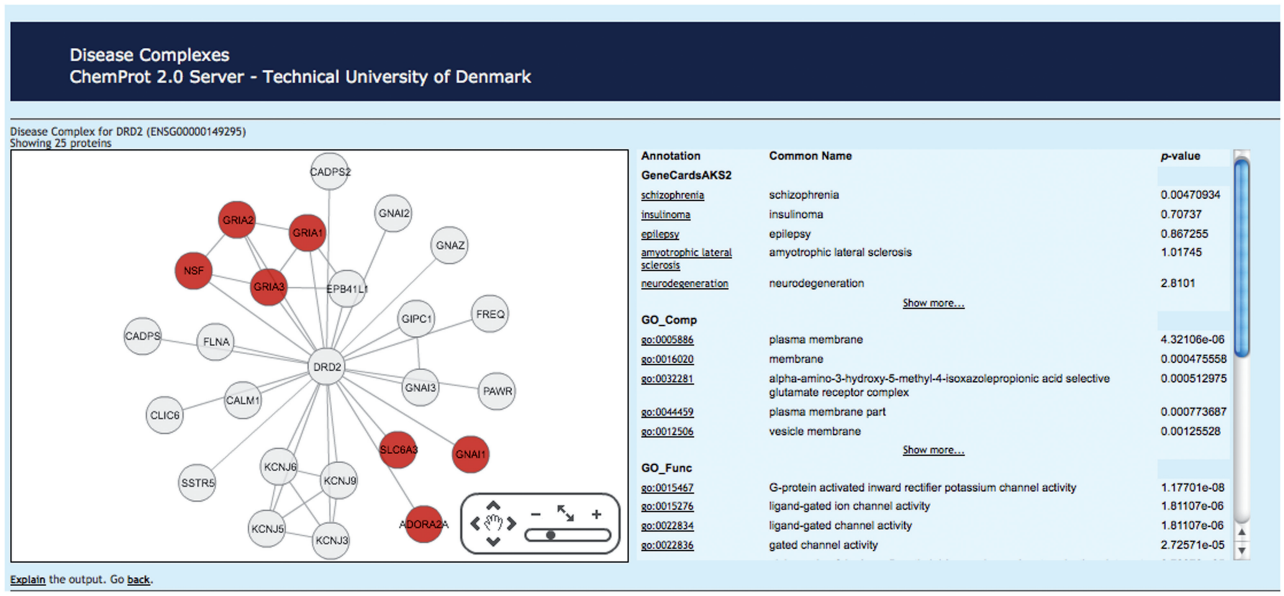

Figure 3. Example of the disease complexes network representation for the dopamine receptor D2 (DRD2). Twenty-five proteins interact directly to the protein DRD2 and pointing the cursor to 'Schizophrenia', seven genes are associated to this disease.

associated to it) active on 470 proteins are displayed. Some of these drugs (ropinirole, pergolide, amantadine and pramipexole) are used for the treatment of Parkinson diseases, by affecting the dopaminergic and serotonergic systems. Interestingly, visual hallucinations are symptoms of the Parkinson's disease and perturbing the serotonergic system could help to alleviate these symptoms (36). Another interesting aspect is that these drugs affect several proteins associated to 'Bone' and osteoporosis disease. For example, there is a possible association between the polymorphism of the serotonin transporter (HTT) and the development of osteoporosis (37). Some of these drugs bind to HTT and could thus be potentially investigated for drug repurposing.

Many diseases seem not to be the result of a single defect but are rather caused by multiple molecular and cellular abnormalities. Therefore, observations of a drug effect not only at the molecular level but also at cellular and systems levels should guide therapeutic strategies for the development of better and safer drugs. ChemProt-2.0 offers the possibility of interrogating multiple layers of information by linking chemically induced biological perturbations to disease and phenotype. We believe with the advances in proteomics, metabolomics and other –omics sciences, combined with next-generation sequencing technologies, we will no longer evaluate the bioactivity profile of a chemical solely at the molecular level, but rather we will investigate biomedical knowledge with the integration of genetic polymorphisms and clinical effects (38).

FUNDING

The Innovative Medicines Initiative Joint Undertaking, OPENPHACTS (to L.W., J.K.); Innovative Medicines Initiative Joint Undertaking, eTOX (to U.P.J., O.T.); GENDINOB project supported by the Danish council for Strategic Research (to K.A.); Lundbeck foundation (to S.K.K.); National Institutes of Health [R21 GM095952 to T.I.O.]. Funding for open access charge: The EU-IMI project OPENPHACTS.

Conflict of interest statement. None declared.

REFERENCES

1. Mullard,A. (2011) Accelerated approval dust begins to settle. *Nat. Rev. Drug. Discov.*, **10**, 797–798.
2. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovitch,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
3. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccion,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
4. Williams,A.J., Ekins,S. and Tkachenko,V. (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discov. Today*, **17**, 685–701.
5. Rognan,D. (2007) Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.*, **152**, 38–52.
6. Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.
7. Oprea,T.I., Tropsha,A., Faulon,J.L. and Rintoul,M.D. (2007) Systems chemical biology. *Nat. Chem. Biol.*, **3**, 447–450.
8. Wild,D.J., Ding,Y., Sheth,A.P., Harland,L., Gifford,E.M. and Lajiness,M.S. (2012) Systems chemical biology and the semantic web: what they mean for the future of drug discovery research. *Drug Discov. Today*, **17**, 469–474.
9. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2012) Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**, 1741–1748.
10. Hansen,N.T., Brunak,S. and Altman,R.B. (2009) Generating genome-scale candidate gene lists for pharmacogenomics. *Clin. Pharmacol. Ther.*, **86**, 183–189.
11. Kuhn,M., Szklarczyk,D., Franceschini,A., von Mering,C., Jensen,L.J. and Bork,P. (2012) STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Res.*, **40**, D876–D880.
12. Bauer-Mehren,A., Rautschka,M., Sanz,F. and Furlong,L.I. (2010) DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics*, **26**, 2924–2926.
13. von Eichborn,J., Murgueitio,M.S., Dunkel,M., Koerner,S., Bourne,P.E. and Preissner,R. (2011) PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.*, **39**, D1060–D1066.
14. Taboureau,O., Nielsen,S.K., Audouze,K., Weinhold,N., Edsgärd,D., Roque,F.S., Kouskoumvekaki,I., Bora,A., Curpan,R., Jensen,T.S. *et al.* (2011) ChemProt: a disease chemical biology database. *Nucleic Acids Res.*, **39**, D367–D372.
15. Lage,K., Karlberg,E.O., Størling,Z.M., Olason,O.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tümer,Z., Pociot,F., Tommerup,N. *et al.* (2007) A human phenome–interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
16. Keiser,M.J., Roth,B.L., Armbruster,B.N., Ernsberger,P., Irwin,J.J. and Shoichet,B.K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
17. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) Binding DB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
18. Roth,B., Lopez,E., Beischel,S., Weskaemper,R.B. and Evans,J.M. (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.*, **102**, 99–110.
19. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
20. McDonagh,E.M., Whirl-Carrillo,M., Garten,Y., Altman,R.B. and Klein,T.E. (2011) From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med.*, **5**, 795–806.
21. Olah,M., Rad,R., Ostopovici,L., Bora,A., Hadaruga,N., Hadaruga,D., Moldovan,R., Fulias,A., Mracec,M. and Oprea,T.I. (2007) WOMBAT and WOMBAT-PK: bioactive databases for lead and drug discovery. In: Schreiber,S.L., Kapoor,T.M. and Wess,G. (eds), *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*. Wiley-VCH, New York, pp. 760–786.
22. Sharman,J.L., Mpamhanga,C.P., Spedding,M., Germain,P., Staels,B., Dacquet,C., Laudet,V., Harmar,A.J. and NC-IUPHAR. (2011) IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.*, **39**, D534–D538.
23. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The comparative Toxicogenomics database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.
24. De Smet,P.A.G.M. (1993) New applications of the ATC/DDD methodology in the Netherlands part 1. ATC/DDD principles and computerized medication surveillance. *Int. Pharm. J.*, **7**, 196–199.
25. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.*, **32**, 564–567.
26. Stelzer,G., Dalah,I., Stein,T.I., Satanower,Y., Rosen,N., Nativ,N., Oz-Levi,D., Olender,T., Belinky,F., Bahir,I. *et al.* (2011) In-silico human genomics with GeneCards. *Hum. Genomics*, **5**, 709–717.
27. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
28. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
29. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
30. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabasi,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

31. Durant,J.L., Leland,B.A., Henry,D.R. and Nourse,J.G. (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280.
32. O'Boyle,N., Banck,M., James,C., Morley,C., Vandermeersch,T. and Hutchison,G. (2011) Open Babel: an open chemical toolbox. *J. Cheminformatics*, **3**, 33.
33. UniProt Consortium. (2011) Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.*, **39**, D214–D219.
34. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids. Res.*, **25**, 3389–3402.
35. Oprea,T.I., Nielsen,S.K., Ursu,O., Yang,J.J., Taboureau,O., Mathias,S.L., Kouskoumvekaki,I., Sklar,L.A. and Bologa,C.G. (2011) Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Mol. Inf.*, **30**, 100–111.
36. Politis,M. and Loane,C. (2011) Serotonergic dysfunction in Parkinson's disease and its relevance to disability. *Scientif. World J.*, **11**, 1726–1734.
37. Ferreira,J.T., Levy,P.Q., Marinho,C.R., Bicho,M.P. and Mascarenhas,M.R. (2011) Association of serotonin transporter gene polymorphism 5HTTVNTR with osteoporosis. *Acta Reumatol. Port.*, **36**, 14–19.
38. Oprea,T.I., Taboureau,O. and Bologa,C.G. (2012) Of possible cheminformatics futures. *J. Comput. Aided Mol. Des.*, **26**, 107–112.